Proceedings of TerraBytes Workshop at the 42nd International Conference on Machine Learning 1–16, 2025

Shaping Fine-Tuning of Geospatial Foundation Models: Effects of Label Availability and Temporal Resolution

Giovanni Castiglioni GIOVANNI.CASTIGLIONI@CENIA.CL
Department of Computer Science, Universidad de Chile, Santiago, Chile
Centro Nacional de Inteligencia Artificial, Macul, Chile

Nicolás Isla
Centro de Información de Recursos Naturales, Santiago, Chile
Department of Computer Science, Universidad de Chile, Santiago, Chile

Cristian B. Calderon
Centro Nacional de Inteligencia Artificial, Macul, Chile

Javiera Castillo-Navarro
CEDRIC, Conservatoire National des Arts et Métiers, Paris, France

Sébastien Lefèvre
IRISA, Université Bretagne Sud, UMR 6074, Vannes, France
UiT – The Arctic University of Norway, Tromsø, Norway

Valentin Barriere
Department of Computer Science, Universidad de Chile, Santiago, Chile
Centro Nacional de Inteligencia Artificial, Macul, Chile

Abstract

Fine-tuning foundation models is a key step in adapting them to a particular task. In the case of Geospatial Foundation Models (GFMs), fine-tuning can be particularly challenging given data scarcity both in terms of the amount of labeled data and, in the case of Satellite Image Time Series (SITS), temporal context. Under these circumstances, the optimal GFM fine-tuning strategy across different labeled data regimes remains poorly understood. In this paper, we thoroughly assess and study the performances of two different GFMs given several combinations of two data scarcity factors: the number of labeled samples and the sequence length. Specifically, we analyze the performances on a crop classification task, particularly, semantic segmentation of the Sentinel-2 images contained in the PASTIS-HD dataset. We compare GFMs to U-TAE, as a fully supervised baseline, across varying amounts of labeled data (1%, 10%, 50%, 100%) and temporal input lengths (1, 6, 15, 25 and 35). Among these explorations, we find that using a smaller learning rate for the pre-trained encoders improves performance in moderate and high data regimes (50%-100%). In contrast, full fine-tuning outperforms partial fine-tuning in very low-label settings (1%-10%). This behavior suggests a nuanced trade-off between feature reuse and adaptation that defies the intuition of standard transfer learning. The code is available here.

Keywords: Foundation Models, Fine-Tuning, Time-Series, Data Scarcity

1. Introduction

Self-supervised learning (SSL) allows the encoding of knowledge into the parameters of so-called foundation models, which through this kind of training leverages large unlabeled datasets to obtain high quality representations (Devlin et al., 2019). These models typically display powerful transfer learning capabilities, reaching higher performances in tasks where the data can be scarce in terms of labeled examples (Yu et al., 2022; Marszalek et al., 2022; Zou et al., 2023). Geospatial Foundation Models (GFMs) are instances of these type of models that have been pre-trained on huge datasets of Satellite Images or Satellite Image Time Series (SITS) (Dumeur, 2024). Processing of remote sensing data, which has traditionally focused on manual interpretation and task-specific models, has recently been revolutionized by the advent of these large-scale and pre-trained new methods (Lu et al., 2024). GFMs and other classical pre-trained models showed competitive performances in various tasks, including Crop Type Mapping (Dumeur et al., 2024a; Chang et al., 2024). They are known to be especially better in low-data and label regimes, which makes them useful as labeled data collection for geospatial applications can be expensive (Rolf et al., 2024).

However, recent work suggests that GFMs perform worse than Task-Specific Models (TSMs) when labeled data are abundant, even though foundation models typically have two to three orders of magnitude more parameters (Marsocci et al., 2024). These results position machine learning for satellite data as a unique testbed, where typical methods and techniques, which resulted beneficial in some modalities, may not necessarily apply to satellite images (Rolf et al., 2024). As such, in this study, we take a deeper dive into this phenomenon, focusing specifically on the task of Crop Type Mapping (Garnot and Landrieu, 2021), auditing two differently complex GFMs, and evaluating whether the performance relation between GFMs and task-specific models is agnostic to the data regime (i.e., small training set and/or short time series). The latter point is motivated by the observation that fine-tuning of large pre-trained models can be instable, particularly in the case of data scarcity (Mosbach et al., 2021; Zhang et al., 2021).

In this study, we systematically compare two GFMs and a TSM (acting as baseline), and first propose to study the effect of the input data sequence length effect, in order to evaluate how much context is needed to reach respectable performances. Second, we propose to study the effect of label scarcity and compare the performance of these models when fine-tuned with parametrically controlled decreasing amounts of data.

We found that GFMs are competitive with TSMs for Crop Type Mapping even in the case of a large fine-tuning dataset. Nonetheless, the length of the input data sequences affects the performance of GFMs and TSMs in a similar way, i.e., accuracy patterns plateau in a logarithmic manner.

Additionally, we find that the generalization capabilities of GFMs crucially depend on the fine-tuning strategy, under particular data regimes. In particular, this strategy consists of weighting the parameter update of the pretrained models, specifically scaling down the learning rate of the encoder. This is done to control the alignment of the latter to the task-specific requirements in terms of feature extraction, manipulating the trade-off between perturbing the already learned features to solve a downstream task, and leveraging those features when the data scarcity conditions or the fine-tuning strategy are not sufficient to accomplish the downstream task.

We list the main contributions of this work as follows:

- Benchmarking two differently complex geospatial foundation models (CROMA (Fuller et al., 2023), SSL4EO's DINO (Stewart et al., 2023)) against a fully-supervised U-TAE baseline (Garnot and Landrieu, 2021) in the PASTIS-HD Sentinel-2 dataset under four label budgets and five temporal resolutions, providing a thorough analysis in real-world data constraints. In this context, we found that even with abundant labels, some fine-tuned models can exhibit a consistent advantage over the task-specific model.
- We propose to use a specific fine-tuning strategy where a hyperparameter controls the learning rate of the encoder relative to the decoder, isolating the effect of encoder adaptation on downstream performance and allowing us to evaluate how much tuning is beneficial under different data regimes.

2. Related Works

Crop Segmentation Garnot and Landrieu (2021) were among the first to propose a large scale dataset (>124k parcels) of crop segmentation for deep neural networks. They introduced the PASTIS dataset, and the U-TAE, a model based on the aggregation of a U-Net and a Temporal Attention Encoder (TAE) in order to adapt the U-Net architecture to SITS. This dataset was augmented a few years later with images from Sentinel-1 into PASTIS-R (Sainte Fare Garnot et al., 2022), and with images from Very High Resolution (Garioud et al., 2023). Rustowicz et al. (2019) also propose to tackle this task, using sparse ground truth labels composing 4 or 5 classes in Sudan and Ghana.

Nowadays, there has been a switch to larger pre-trained SSL models (Yuan and Lin, 2021). Dumeur et al. (2024b,a); Dumeur (2024) proposed several architectures using SSL methods applied to crop segmentation, such as the Unet-BERT spAtio-temporal Representation eNcoder (U-BARN) to exploit irregularly sampled SITS. They use dense sequences (up to 100 timesteps) to classify the pixels.

Recently, Reuss et al. (2025) proposed to study few-shot time series classification with basic transformers using the EuroCropsML dataset between Portugal, Estonia and Latvia (Reuss et al., 2024). Barriere et al. (2024); Barriere and Claverie (2022) consider the same crop taxonomy to study few-shot learning between France and the Netherlands. Both highlighted the importance of domain adaption or pre-training of the model on domain data.

Foundation Models Many works proposing GFMs have been published in the last months (see Lu et al. (2024) for a review of remote sensing foundation models). They rely on training SSL models on huge datasets (Nedungadi et al., 2024; Guo et al., 2023; Bastani et al., 2023).

Prithvi and Prithvi2.0 (Jakubik et al., 2023; Szwarcman et al., 2024) are the largest publicly available models at the time, with 600M parameters. Fuller et al. (2023) propose Contrastive Radar-Optical Masked Autoencoders (CROMA), which jointly learn two modalities (Radar and Optical) using both contrastive and masked reconstruction losses. Copernicus-FM (Wang et al., 2025) is a model fully dedicated to Copernicus data, such as Sentinel-1, Sentinel-2, or Sentinel-5. Still with Sentinel data, series of datasets called SSL4EO-S12 have been released (Wang et al., 2023; Blumenstiel et al., 2025) and used to train models based on SSL architectures like DINO (Caron et al., 2021), MAE (He et al., 2022) and MoCo (Chen and Xie, 2021). Moreover, Vision Transformers (ViTs) have also been used succesfully. For instance, Tarasiou et al. (2023) propose a ViT that splits satellite image time series into temporo-then-spatial patches, uses date-aware positional encodings and multiple class tokens, and processes them with factorized attention. Their model outperforms CNN/RNN baselines in crop-type segmentation and classification tasks by wide margins with similar model size and inference time. Similarly, Bountos et al. (2023) propose FoMo-Bench, a unified forest monitoring benchmark, useful to assess the ability of GFMs. Furthermore, these authors propose FoMo-Net, a sensor-agnostic ViT pretrained to fuse optical, SAR, LiDAR and other bands across scales, achieving strong performance on zero-/few-shot classification, segmentation and detection tasks. Yet, no RSFM has exhibited universal superiority across all downstream tasks so far (Adorni et al., 2025).

GFMs for Crop Segmentation Regarding agriculture applications of GFMs, Chang et al. (2024) are studying the generalizability of GFM for Crop Type Mapping and proposing the Crop Type Bench. They compare SSL4EO-S12 and SatlasPretrain on a benchmark composed of several datasets for crop segmentation, however they do not consider the time series while temporal information is of paramount importance for distinguishing between crop types.

AnySat (Astruc et al., 2024b) and OmniSat (Astruc et al., 2024a) are two SSL methods focusing on jointly learning multimodal representation by exploiting the alignments of the modalities. The networks remain performant at inference phase with one modality only. AnySat, based on joint embedding predictive architecture (JEPA; Assran et al. 2023), obtains state-of-the-art results on the entire PASTIS dataset.

Galileo (Global and Local Flexible Earth Observation models; Tseng et al. 2025) is a GFM also evaluated on the PASTIS dataset, without fully fine-tuning the model but only doing linear probing. Guo et al. (2023) propose SkySense, a GFM that jointly learns from time-series optical (RGB + multispectral) and synthetic aperture radar (SAR) data. Their model was trained on millions of spatiotemporal sequences, and uses a factorized spatiotemporal encoder, multi-granularity contrastive learning, and geo-context prototypes to create transferable pixel-, object- and image-level features. Evaluated on distinct tasks (e.g., segmentation, detection, change detection, crop mapping), SkySense outperformed several prior remote-sensing foundation models.

Nedungadi et al. (2024) propose MMEarth, a multi-modal global dataset used to train a multi-pretext masked autoencoder that reconstructs diverse pixel- and image-level signals to learn representations for Sentinel-2 imagery. Their model outperforms models pretrained on ImageNet and on single-modality satellite-images on several land-cover classification and segmentation benchmarks, especially in low-label settings.

Of particular relevance to our work, Marsocci et al. (2024) introduce PANGAEA, a globally diverse benchmark that spans multiple domains, sensor modalities, resolutions and temporalities to standardize GFM evaluation. Their result suggest that current GFMs, although versatile, often fail to consistently outperform simpler supervised baselines, especially when their pre-training data poorly match downstream tasks distribution (Rolf et al., 2021), highlighting the need for more robust multi-modal, multi-temporal pre-training.

3. Methodology

3.1. Models

In this work, we considered two different GFMs based on their representativeness of the existing SSL families (Balestriero et al., 2023), model complexity (see Tab. 1) and promising previous results shown by Marsocci et al. (2024), and a state-of-the-art model on the relevant downstream task acting as a fully-supervised specialized network:

- CROMA (Fuller et al., 2023): A GFM pre-trained via masked auto-encoding and contrastive learning on 3M patches of multispectral Sentinel-2 data.
- SSL4eo-DINO (Blumenstiel et al., 2025): A GFM pre-trained with self-supervised learning on 3M patches of EO imagery using DINO, a self-supervised method based on auto-distilled representations.
- U-TAE (Garnot and Landrieu, 2021): A supervised baseline with temporal attention, trained from scratch on each data regime.

Table 1: Model sizes in terms of Trainable Parameters for two configurations: frozen encoder and whole network.

Model	# of Trainable Parameters (M)	
	Only Decoder	Whole network
CROMA	46.95	350.0
DINO	30.89	53.5
U-TAE	-	1.1

In order to compare the models fairly and adapt them to SITS, we conducted the experiments for both pretrained models using the same decoder architecture, which is an UPerNet (Xiao et al., 2018), and performing temporal aggregation using a Time-Attention Encoding module to aggregate w.r.t time, following the same methodology adopted by Marsocci et al. (2024).

FT-Rate We control the adaptation of pre-trained encoders via a hyperparameter defined as the FT-Rate, which scales the encoder's learning rate relative to the untrained decoder's. This means that FT-Rate = 0.0 freezes the encoder, FT-Rate = 0.1 applies a 10x smaller learning rate, and FT-Rate = 1.0 uses the same learning rate for both the encoder and the decoder.

3.2. Data Scarcity

To assess the fine-tuning of GFMs, we conducted experiments on the PASTIS-HD dataset based on multi-temporal Sentinel-2 imagery. The dataset comprises 2,433 agricultural parcels in France with pixel-level annotations.

To simulate varying levels of supervision, we sub-sample the training set into four label regimes: 1%, 10%, 50%, and 100% of the available labeled parcels in the training set. To do this, similar to the stratified methodology adopted by Marsocci et al. (2024), we generate bins for each patch of the entire set, based on the amount of presence of a class within an image at a pixel-level. This results in a quantized histogram representing a coarse distribution of the classes.

To preserve a distribution similar to that of the original training set, we compute the average quantized histogram of the dataset and select the desired percentage of samples with the smallest Jensen-Shannon divergence (JSD). The JSD is a symmetric and smoothed version of the Kullback–Leibler (KL) divergence and is defined as:

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2}\,\mathrm{KL}(P \parallel M) + \frac{1}{2}\,\mathrm{KL}(Q \parallel M),$$

where P is the average quantized histogram, Q is the quantized histogram of a patch, and $M = \frac{1}{2}(P + Q)$. The resulting distributions of this process, relative to each subset of the original training set, are shown in Fig. 1.

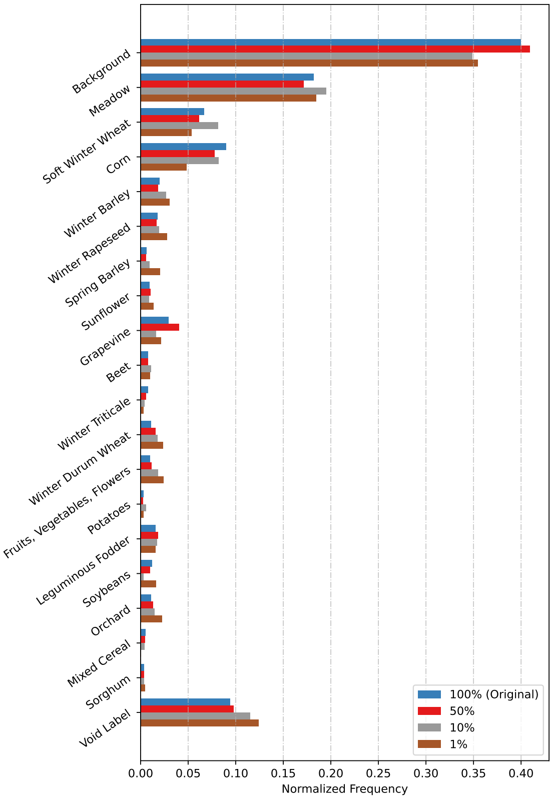


Figure 1: Pixel-level distributions of the training set for each class, in the 4 different adopted label-scarce regimes (100%, 50%, 10%, 1%).

This method differs from the one of Marsocci et al. (2024), which originally led to a high quantity of samples selected, even though the selected percentage of data was small, because at least one sample of each generated bin was included in the subset. With our proposed methodology, it is not necessary to include one sample of each bin, but only the desired percentage of samples that are most similar to the original distribution of the training set.

3.3. Limited Sequence Length

Five temporal depths (1, 6, 15, 25, 35 dates) were considered to test the capacity of pretrained models to capture phenological patterns. To simulate limited sequence length, we select 35 time instances, as evenly-spaced as possible, for each patch, and then generate nested subsets from those 35 selected time instances. Particularly, we consider subsets of sizes 1, 6, 15, 25 and 35. The subset with 1 temporal acquisition contains only the last available instance.

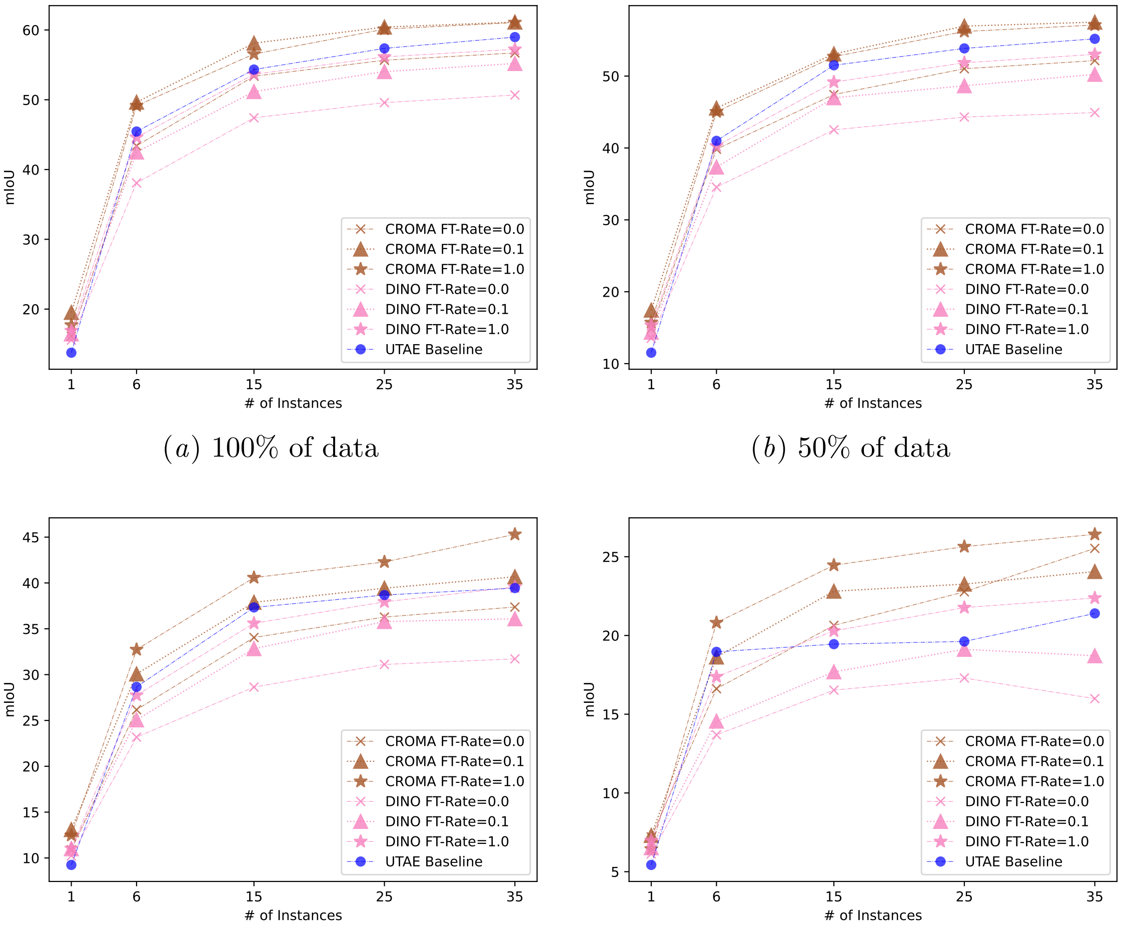
In order to ensure that the quantity of information contained in the smallest sequences is also contained in the largest ones, we create them in a nested way. Each subset, composed of samples extracted from the 35 originally selected instances, is defined by time instances as evenly-spaced as possible from the immediately larger set, i.e., 25 instances are selected from the original 35, then 15 instances are selected from those 25, and then 6 instances are selected from those 15, ensuring that every subset is contained in the larger ones in a nested manner that preserves temporal context.

4. Experiments and Results

4.1. Label Availability

Performance is classically evaluated on the basis of the achieved mean Intersection over Union (mIoU). We tested data-scarce scenarios, the first one being the least aggressive, with 50% of the original training set included, while the validation and test sets remain the same. In this case, Table 2 (right part) and Figure 2(b) show patterns similar to those appreciated in Table 2 (left part) and Figure 2(a) where 100% of the data were used, with partial fine-tuning marginally and consistently outperforming other strategies for the larger encoder, while U-TAE remained above all the SSL4eo-DINO experiments using multi-temporal data.

The second data-scarce configuration explored was more aggressive, with 10% of the training data being considered for all mod-



(a) 100% of data

(b) 50% of data

(c) 10% of data

(d) 1% of data

Figure 2: mIoU per number of instances for models trained on different levels of data scarcity. Each line corresponds to a specific model and encoder fine-tuning rate (FT-Rate), 0.0 corresponding to a fully frozen encoder. U-TAE is shown as a fully-supervised baseline without encoder pretraining.

els. Table 3 (left part) and Figure 2(c) show a change in the relation between partial fine-tuning and full fine-tuning, with the latter outperforming the former, and both being the best compared to the rest of the configurations. In this setting, the least complex model, SSL4eo-DINO, manages to reach the performance of U-TAE when fully fine-tuned using 35 instances, which did not happen in previous experiments, starting to exhibit the benefit that even small GFMs have over fully-supervised methods under label-scarcity.

Finally, we trained with 1% of the original training data, making this configuration the most limited in terms of labels available for the model to learn to segment crop fields. Table 3 (right part) and Figure 2(d) show the advantage that GFMs have over the fully-supervised model.

Table 2: mIoU per instance for CROMA, SSL4eo-DINO, and U-TAE trained with 100% and 50% of the Sentinel-2 labeled samples from the PASTIS dataset. FT-Rate denotes the factor applied to scale the encoder's learning rate relative to the decoder. The best result for each multi-temporal configuration is highlighted in **bold**.

Model	FT-Rate	100% of data					50% of data				
		1	6	15	25	35	1	6	15	25	35
CROMA	0.0	16.10	43.40	53.37	55.66	56.71	14.83	39.83	47.43	51.02	52.15
	0.1	**19.52**	**49.65**	**58.11**	**60.41**	**61.12**	**17.44**	**45.54**	**53.08**	**56.95**	**57.50**
	1.0	17.71	49.15	56.55	60.12	61.06	15.71	45.03	52.71	56.20	57.10
DINO	0.0	15.55	38.08	47.41	49.59	50.68	13.52	34.55	42.52	44.30	44.92
	0.1	16.43	42.50	51.17	54.03	55.19	14.36	37.34	46.98	48.65	50.24
	1.0	16.91	44.61	53.65	56.14	57.23	15.34	40.28	49.16	51.85	53.02
U-TAE	-	13.75	45.45	54.33	57.36	58.98	11.52	40.99	51.50	53.85	55.18

Using this configuration, when it is fully fine-tuned, SSL4eo-DINO consistently outperforms U-TAE, which stalls its performance with the increasing number of instances. The latter is also notably surpassed by some settings of frozen CROMA encoders given sufficiently long time sequences. Moreover, these frozen encoders interestingly surpass the performance of some partially fine-tuned CROMA models.

In both the 1% and 10% label regimes, CROMA and SSL4eo-DINO significantly outperform U-TAE, even when the encoder is frozen in some cases, underscoring the benefit of large-scale pretraining when supervision is limited.

4.2. Temporal Resolution

Across all models and label regimes, performance improves consistently and plateaus with the increment of the number of temporal observations. The greatest gains occur in low-label settings, highlighting the value of phenological information when supervision is scarce. This can be seen in Figure 2, which shows a logarithmic behavior on the progression of mIoU across an increasing number of instances on all the regimes of labeled data.

We compare U-TAE along with CROMA and SSL4eo-DINO on the three fine-tuning configurations. For settings with both 100% and with 50% of training data, U-TAE is the second best model, outperforming every configuration of the SSL4eo-DINO encoder with considerably less parameter complexity, except for the one using mono-instance patches; as expected, given that the main feature of U-TAE lies on its Time-Attention Encoder module.

In the middle example of Figure 4, the model trained on the entire dataset succeeds to detect sorghum (pink) when using at least 15 instances, highlighting the need of temporal context to detect classes at the tail of the distribution.

4.3. Learning Rate Policy

For the first case, Table 2 shows the mIoU per number of instances for the models and the fine-tuning strategies using 100% of the original training set. FT-Rate indicates the scaling factor applied to the encoder's learning rate relative to the decoder. The results show that using a moderate fine-tuning strategy (i.e., FT-Rate = 0.1), although marginally, consistently improves the performance of CROMA when the number of instances increases, achieving the highest mIoU in all instances counts. In this sense, partial fine-tuning (FT-Rate=0.1) achieves

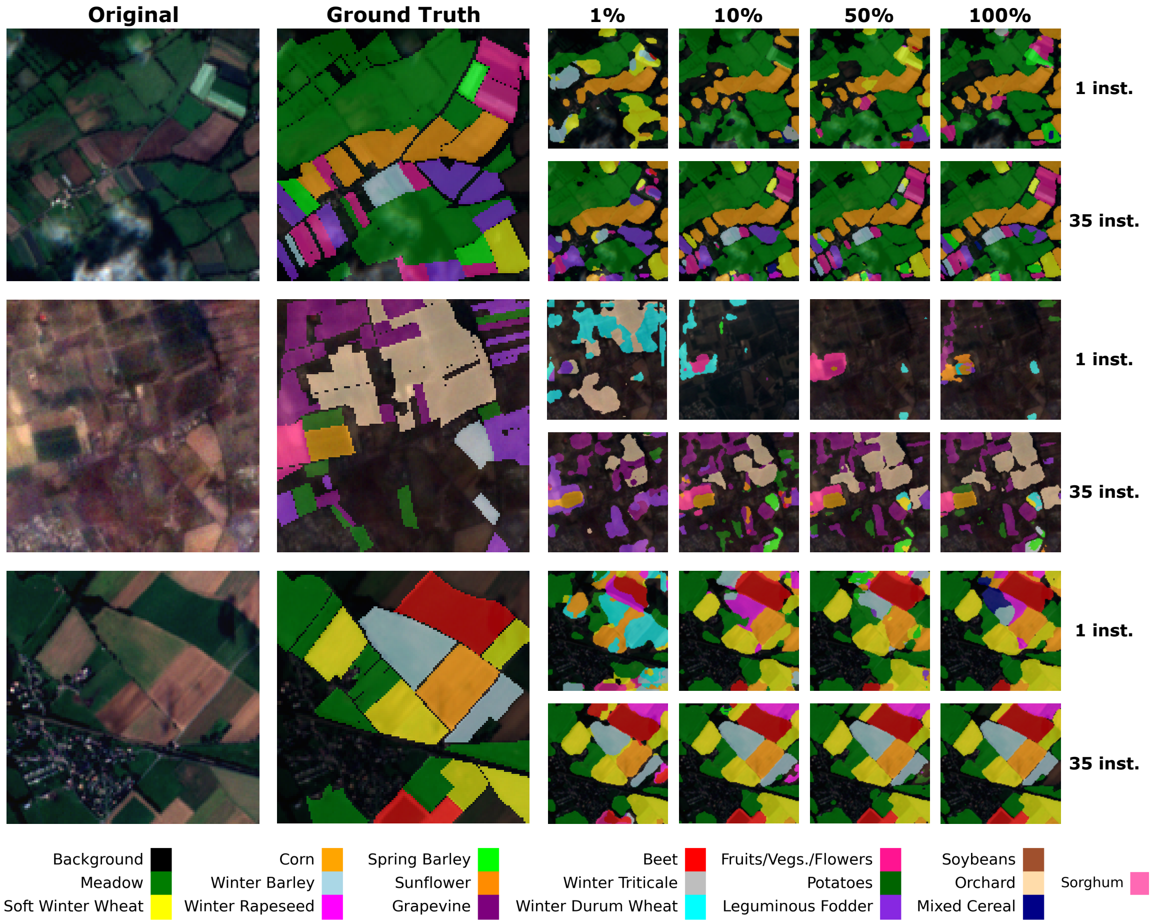


Figure 3: Segmentations produced for three different images on two levels of temporal resolution in all the configurations of label scarcity. **Left to right** shows the original image, the ground truth label, and two rows of increasing label availability regimes, the **top row** shows results after training with 1 instance per patch, and the **bottom row** shows results after training with 35 instances per patch.

superior performance for CROMA, the larger model, highlighting the benefit of gentle adaptation when ample labeled data are available for large encoders. This pattern is also observed in Table 2 (right part) and Figure 2(*b*) for regimes using 50% of the labels. These findings suggest a non-monotonic relationship between label availability and optimal FT-rate. With very few samples, full fine-tuning (FT-Rate = 1.0) is beneficial. With moderate or ample data, conservative fine-tuning (FT-Rate = 0.1) preserves useful representations. Freezing encoders (FT-Rate = 0.0) offers a strong baseline, but usually underperforms partial or full adaptation. These dynamics challenge the intuition from natural image transfer learning, highlighting the need for domain-specific strategies in remote sensing. Several factors may contribute to this unexpected result, e.g., with few labels a small learning rate may not sufficiently adapt neither the encoder, nor the

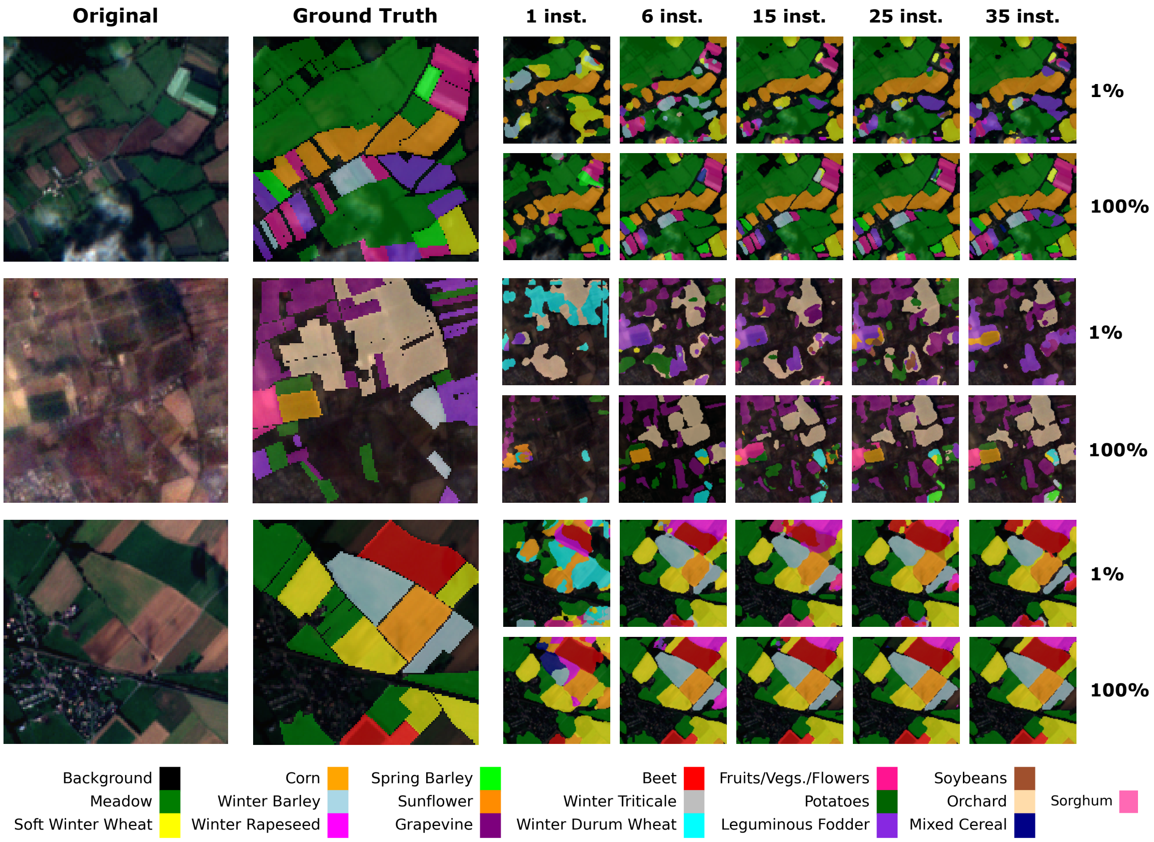
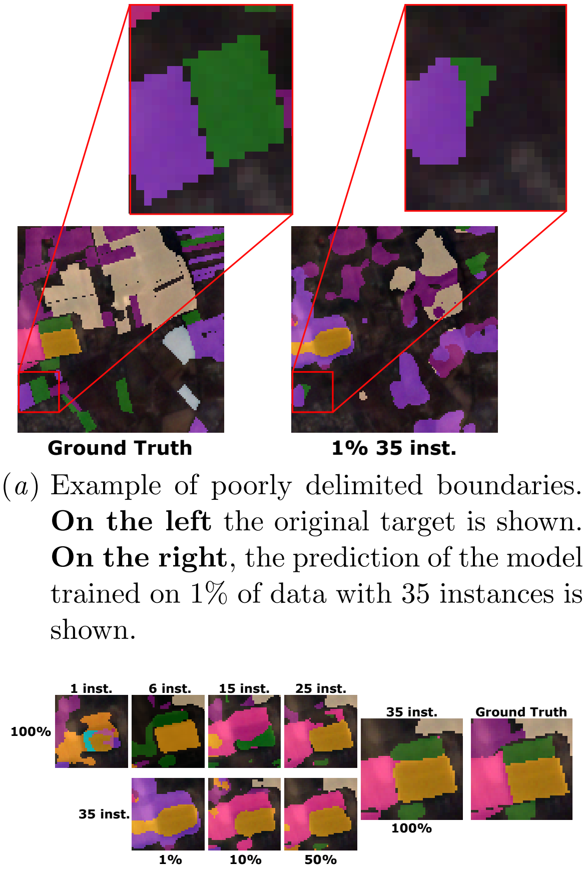


Figure 4: Segmentations produced for three different images on two levels of label availability in all the configurations of temporal resolutions. **Left to right** shows the original image, the ground truth label, and two rows of increasing temporal resolution regimes, the **top row** shows results after training with 1% of the data, and the **bottom row** shows results after training with 100% of the data.

Table 3: mIoU per instance for CROMA, SSL4eo-DINO, and U-TAE trained with 10% and 1% of the Sentinel-2 labeled samples from the PASTIS dataset. FT-Rate denotes the factor applied to scale the encoder's learning rate relative to the decoder. The best result for each multi-temporal configuration is highlighted in **bold**.

Model	FT-Rate	10% of data					1% of data				
		1	6	15	25	35	1	6	15	25	35
CROMA	0.0	10.89	26.17	34.05	36.28	37.37	7.21	16.62	20.64	22.77	25.53
	0.1	**13.09**	30.07	37.89	39.43	40.67	**7.30**	18.63	22.81	23.26	24.06
	1.0	12.47	**32.73**	**40.58**	**42.29**	**45.30**	6.44	**20.81**	**24.46**	**25.64**	**26.42**
DINO	0.0	10.23	23.18	28.64	31.11	31.72	6.13	13.69	16.53	17.30	15.99
	0.1	11.00	25.06	32.83	35.79	36.09	6.52	14.56	17.69	19.12	18.71
	1.0	11.03	27.75	35.60	37.94	39.56	6.98	17.38	20.29	21.77	22.38
U-TAE	-	9.24	28.66	37.33	38.68	39.44	5.43	18.96	19.45	19.62	21.40

randomly initialized decoder, which cannot compensate for features misaligned with the task. Furthermore, even pretrained representations may not align well with segmentation at high spatial resolution, so aggressive adaptation may help to correct this misalignment. These results also suggest that FT-Rate should be treated as a relevant hyperparameter and not fixed a priori.



(*a*) Example of poorly delimited boundaries. **On the left** the original target is shown. **On the right**, the prediction of the model trained on 1% of data with 35 instances is shown.

(*b*) Progression of models under temporal scarcity (**on the top**) and label scarcity (**on the bottom**) converging into the ground truth target.

Figure 5: Detailed comparison of ground truth and segmentations produced by data-scarce models. (a) portrays the poor definition of boundaries, even with correct classification. (b) shows the progressive improvement of the segmentations when enriching the data quality.

4.4. Semantic Map Visualization

An overall visualization of the effects of varying the temporal resolution and the availability of labels on the segmentations that the models produce is shown in Figure 3 and Figure 4. Scarce regimes have less defined regions and struggle with the correct delineation of crop fields, generating rounded edges and poorly classified regions, even though the label can be correct, as shown in Figure 5.

Conversely, richer configurations are more able to capture the actual parcellation and depiction of the landscape. Indeed, when comparing the triple parcels of sorghum, meadow and corn in Figure 5(*b*), we can see on the one hand the 100% model needs the full data sequence to classify it perfectly. On the other hand, getting the full data sequence is not enough as the 35 instance models trained with 50% of the data still strugle to get the three parcels right.

5. Conclusion

Our findings underscore that fine-tuning is crucial for both GFMs in the studied dense segmentation task. Pre-trained GFMs without adaptation performed poorly, whereas fine-tuned GFMs achieved large gains (e.g., CROMA's reported mIoU boost). This contribution stresses that model adaptation must be a standard practice, in line with benchmarks that treat fine-tuning as essential.

GFMs significantly outperform supervised baselines in low-supervision regimes, and the choice of fine-tuning rate plays a relevant role in achieving the strongest performance on the selected dataset. Surprisingly, full fine-tuning is more effective than partial adaptation when labels are extremely scarce, contradicting common transfer learning assumptions about the quality of the pretrained encoder's features. Our results underscore the need to tailor fine-tuning strategies to data availability and task alignment, and to carefully tune FT-Rate as a hyperparameter in fine-tuning experiments.

By contextualizing our results within recent literature (Marsocci et al., 2024), we provide clear recommendations: For large GFMs, tune FT-rate as a key hyperparameter, as no guarantee of alignment of the pretraining and downstream tasks was observed in our experiments; for experiments with sufficiently abundant labels, fully-supervised models with less computational cost were competitive or superior to larger pretrained models, so consider them as a relevant alternative; leverage long temporal sequences, particularly when supervision is limited, because even frozen GFMs substantially benefitted from richer temporal context.

Acknowledgements

G.C., N.I., C.B.C. and V.B. have been funded by the grant Tecnologías Avanzadas TA24I10016 ANID & G.C. by the grant National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning for Earth observation and land cover mapping. Our study contributes to the understanding of GFMs and fine-tuning strategies in label-scarce regimes, which is highly relevant in real-world scenarios where high-quality annotations are expensive or difficult to obtain. Potential societal benefits include improved land use monitoring, agricultural planning, and environmental conservation, particularly in under-resourced regions. We believe there are no foreseeable negative societal consequences or ethical concerns arising from this work.

References

Pierre Adorni, Minh-Tan Pham, Stéphane May, and Sébastien Lefèvre. Towards Efficient Benchmarking of Foundation Models in Remote Sensing: A Capabilities Encoding Approach. *MORSE Workshop at the Conference on Computer Vision and Pattern Recognition 2025*, 5 2025. URL http://arxiv.org/abs/2505.03299.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *CVPR*, pages 15619–15629, 2023. ISSN 10636919. doi: 10.1109/cvpr52729.2023.01499.

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. OmniSat: Self-Supervised Modality Fusion for Earth Observation. In *ECCV*, 2024a. ISBN 9783031733895. doi: 10.1007/978-3-031-73390-1{_}24. URL http://arxiv.org/abs/2404.08351.

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: An Earth Observation Model for Any Resolutions, Scales, and Modalities. 2024b. URL http://arxiv.org/abs/2412.14123.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A Cookbook of Self-Supervised Learning. 2023. URL http://arxiv.org/abs/2304.12210.

Valentin Barriere and Martin Claverie. Multimodal Crop Type Classification Fusing Multi-Spectral Satellite Time Series with Farmers Crop Rotations and Local Crop Distribution. In *Proceedings of 2nd Workshop on Complex Data Challenges in Earth Observation, IJCAI*, volume 3207, pages 50–57, 2022.

Valentin Barriere, Martin Claverie, Maja Schneider, Guido Lemoine, and Raphaël D'Andrimont. Boosting Crop Classification by Hierarchically Fusing Satellite, Rotational, and Contextual Data. *Remote Sensing of Environment*, 2024.

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding. *Proceedings of the IEEE International Conference on Computer Vision*, pages 16726–16736, 2023. ISSN 15505499. doi: 10.1109/ICCV51070.2023.01538.

Benedikt Blumenstiel, Nassim Ait Ali Braham, Conrad M Albrecht, Stefano Maurogiovanni, and Paolo Fraccaro. SSL4EO-S12 v1.1: A Multimodal, Multiseasonal Dataset for Pretraining, Updated. pages 10–13, 2025. URL http://arxiv.org/abs/2503.00168.

Nikolaos Ioannis Bountos, Arthur Ouaknine, and David Rolnick. FoMo-Bench: a multimodal, multi-scale and multi-task Forest Monitoring Benchmark for remote sensing foundation models. 2023. URL http://arxiv.org/abs/2312.10114.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. pages 9650–9660, 2021. doi: 10.1109/iccv48922.2021.00951. URL http://arxiv.org/abs/2104.14294.

Yi-Chia Chang, Adam J. Stewart, Favyen Bastani, Piper Wolters, Shreya Kannan, George R. Huber, Jingtong Wang, and Arindam Banerjee. On the Generalizability of Foundation Models for Crop Type Mapping. 2024. URL http://arxiv.org/abs/2409.09451.

Xinlei Chen and Saining Xie. An Empirical Study of Training Self-Supervised Vision Transformers Xinlei. In *ICCV*, pages 9640–9649, 2021. URL https://github.com/facebookresearch/moco-v3.

Jacob Devlin, Ming-wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019.

Iris Dumeur. *Ouvrir la voie vers des modèles de fondation exploitant les séries temporelles d'images satellites pour le suivi des surfaces continentales.* PhD thesis, 2024.

Iris Dumeur, Silvia Valero, and Jordi Inglada. Paving the way toward foundation models for irregular and unaligned Satellite Image Time Series. pages 1–13, 2024a. URL http://arxiv.org/abs/2407.08448.

Iris Dumeur, Silvia Valero, and Jordi Inglada. Self-Supervised Spatio-Temporal Representation Learning of Satellite Image Time Series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:4350–4367, 2024b. ISSN 21511535. doi: 10.1109/JSTARS.2024.3358066.

Anthony Fuller, Koreen Millard, and James R. Green. CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders. *Advances in Neural Information Processing Systems*, 36(NeurIPS):1–33, 2023. ISSN 10495258.

Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, and Boris Wattrelos. FLAIR: a Country-Scale Land Cover Semantic Segmentation Dataset From Multi-Source Optical Imagery. (NeurIPS):1–27, 2023. URL http://arxiv.org/abs/2310.13336.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks. *Proceedings of the IEEE International Conference on Computer Vision*, pages 4852–4861, 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.00483.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery. In *CVPR*, 2023. ISBN 9798350353006. doi: 10.1109/CVPR52733.2024.02613. URL http://arxiv.org/abs/2312.10115.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June: 15979–15988, 2022. ISSN 10636919. doi: 10.1109/CVPR52688.2022.01553.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi (Steve) Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation Models for Generalist Geospatial Artificial Intelligence. 2023. ISSN 14764687. doi: 10.1038/s41586-023-05881-4.

Siqi Lu, Junlin Guo, James R Zimmer-Dauphinee, Jordan M Nieusma, Xiao Wang, Parker VanValkenburgh, Steven A Wernke, and Yuankai Huo. AI Foundation Models in Remote Sensing: A Survey. pages 1–21, 2024. URL http://arxiv.org/abs/2408.03464.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models. pages 1–46, 2024. URL http://arxiv.org/abs/2412.04204.

Michael L. Marszalek, Bertrand Le Saux, Pierre-Philippe Mathieu, Artur Nowakowski, and Daniel Springer. Self-supervised learning – A way to minimize time and effort for precision agriculture? 2022. URL http://arxiv.org/abs/2204.02100.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the Stability of Fine-Tuning Bert: Misconceptions, Explanations, and Strong Baselines. *ICLR 2021 - 9th International Conference on Learning Representations*, 2021.

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. MMEarth: Exploring Multi-Modal Pretext Tasks For Geospatial Representation Learning. 2024. ISSN 16113349. doi: 10.1007/978-3-031-73039-9{_}10. URL http://arxiv.org/abs/2405.02771.

Joana Reuss, Jan Macdonald, Simon Becker, Lorenz Richter, and Marco Körner. EuroCropsML: A Time Series Benchmark Dataset For Few-Shot Crop Type Classification. pages 1–5, 2024. URL http://arxiv.org/abs/2407.17458.

Joana Reuss, Jan Macdonald, Simon Becker, Konrad Schultka, Lorenz Richter, and Marco Körner. Meta-learning For Few-Shot Time Series Crop Type Classification: A Benchmark On The EuroCropsML Dataset. 2025. doi: 10.1038/s41597-025-04952-7. URL http://arxiv.org/abs/2407.17458.

Esther Rolf, Theodora Worledge, Benjamin Recht, and Michael I. Jordan. Representation Matters: Assessing the Importance of Subgroup Allocations in Training Data. *Proceedings of Machine Learning Research*, 139:9040–9051, 2021. ISSN 26403498.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: Mission Critical - Satellite Data is a Distinct Modality in Machine Learning. *Proceedings of Machine Learning Research*, 235:42691–42706, 2024. ISSN 26403498.

Rose Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. *CVPR Workshops*, 1:75–82, 2019. URL https://github.

Vivien Sainte Fare Garnot, Loic Landrieu, and Nesrine Chehata. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187:294–305, 2022. ISSN 09242716. doi: 10.1016/j.isprsjprs.2022.03.012.

Adam J. Stewart, Nils Lehmann, Isaac A. Corley, Yi Wang, Yi Chia Chang, Nassim Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. SSL4EO-L: Datasets and Foundation Models for Landsat Imagery. *Advances in Neural Information Processing Systems*, 36(NeurIPS):1–21, 2023. ISSN 10495258.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Ankur Kumar, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Trevor Keenan, Paulo Arevalo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Christopher Hain, Robert Kennedy, Bianca Zadrozny, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications. pages 1–33, 2024. URL http://arxiv.org/abs/2412.02732.

Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. ViTs for SITS: Vision Transformers for Satellite Image Time Series. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2023-June:10418–10428, 2023. ISSN 10636919. doi: 10.1109/CVPR52729.2023.01004.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R. Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning Global and Local Features in Pretrained Remote Sensing Models. 2025. URL http://arxiv.org/abs/2502.09356.

Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023. ISSN 21686831. doi: 10.1109/MGRS.2023.3281651.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J. Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiao Xiang Zhu. Towards a Unified Copernicus Foundation Model for Earth Vision. 2025. URL http://arxiv.org/abs/2503.11849.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. In *ECCV*, 2018. doi: 10.1145/2744769.2744912.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models. (1):1–19, 2022. URL http://arxiv.org/abs/2205.01917.

Yuan Yuan and Lei Lin. Self-Supervised Pretraining of Transformers for Satellite Image Time Series Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14 (January):474–487, 2021. ISSN 21511535. doi: 10.1109/JSTARS.2020.3036602.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. Revisiting Few-Sample Bert Fine-Tuning. *ICLR 2021 - 9th International Conference on Learning Representations*, pages 1–22, 2021.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment Everything Everywhere All at Once. 2023. URL http://arxiv.org/abs/2304.06718.